Final project

Car Sales Behavior Across Different Regions

By Jose De Leon

Table of Content

Introduction 3

Table of Versions 3

Purpose of the analysis 3

Project objective 4

Technological tools used 4

Database 4

- Description of the database 4
- Database segmentation 4
- Entity-Relationship Diagram 5
- List of Tables and Their Columns 5
- Transformations performed 8

Visualizations 9

- Objetive 9
- Scope 9
- Calculated Measures 9
- Dashboard tabs 10
- Entity-Relationship Diagram in Power BI 15
- Futures Initiatives 16

Conclusions 16

Introduction

Data analysis is a fundamental tool for understanding patterns, trends, and consumer behavior. By thoroughly analyzing and understanding the database you are working with, you can develop effective strategies to better capture customer attention and, as a result, increase business margins and sales.

Tools like Power BI allow us to visualize large datasets that would otherwise be difficult to interpret, making complex numerical information more accessible and easier to read.

In this project, our goal is not only to conduct a data analysis but also to develop commercial strategies aimed at increasing customer acquisition, improving the company's market position, boosting car sales, and reducing vehicle inventory.

Table of Versions

Version	Cambios	Fecha
Version 1.0	• Description of the dataset topic • Purpose of the analysis	16-4-2024
Version 2.0	• Project objectives • Scope • Entity-relationship diagram • List of tables and their relationships	1-5-2024
Version 3.0	• Transformations applied to the tables • Calculated measures	4-6-2024
Version 4.0	• Version control table • Implemented technological tools • Description of the database used • Images of report sections • Potential future initiatives	22-6-2024
Version 5.0	• Translation to English	10-5-2025

Table 1: Table of versions

Purpose of the analysis

The goal of this analysis is to understand whether all car brands sell equally across regions and socio-economic groups. To explore this, we will conduct a market analysis of sales trends by brand and region using the provided dataset. This will help determine whether sales behavior is consistent or varies by demographic.

This analysis will offer insights into customer preferences based on demographics, enabling the development of more effective regional sales strategies.

Project objective

The objective of this project is to build a **data visualization tool (dashboard)** using a car sales dataset. This dashboard aims to provide end users with a clear and accessible understanding of the dataset's information to support better decision-making.

Additionally, the project seeks to evaluate the initial hypothesis, enabling decision-makers to define their sales strategy based on the results obtained.

Technological tools used

The following tools were implemented throughout the project:

- **PowerPoint** – Used to design custom backgrounds for dashboards.
- **Miro** – Used to create the entity-relationship diagram (ERD).
- **Microsoft Excel** – Used to transform and organize the original dataset into separate entities.
- **Power BI Desktop** – Used to develop the dashboard and create metrics/visuals.
- **Kaggle (https://www.kaggle.com/)** – Source of the original dataset.

Database

The dataset contains **18 columns** and approximately **23,906 records**. It was sourced from Kaggle.

Description of the database

This project focuses on analyzing regional car sales behavior, with the goal of understanding customer preferences in terms of vehicle model, brand, color, and style, depending on their gender and annual income.

In addition, the dataset includes purchase dates and other important factors that influence consumer decisions, such as engine type and transmission type.

Through this analysis, we aim to uncover strategic market insights and identify demographic purchasing patterns, helping stakeholders better understand customer behavior based on different customer profiles.

Database segmentation

The original dataset contained several columns that could be grouped logically. Therefore, the information was reorganized based on related data segments. The original database included information about customers, sales, dealers, and vehicles.

Once the different entities were identified, the decision was made to separate the original dataset into distinct entities.

The entities are as follows:

- **Customer entity** (Customers)

- **Sales entity** (Sales)
- **Dealer entity** (Dealers)
- **Vehicle entity** (Vehicles)

These entities were later used to build the entity-relationship model.

Entity-Relationship Diagram

Below is the entity-relationship diagram (ERD) for the car sales dataset.

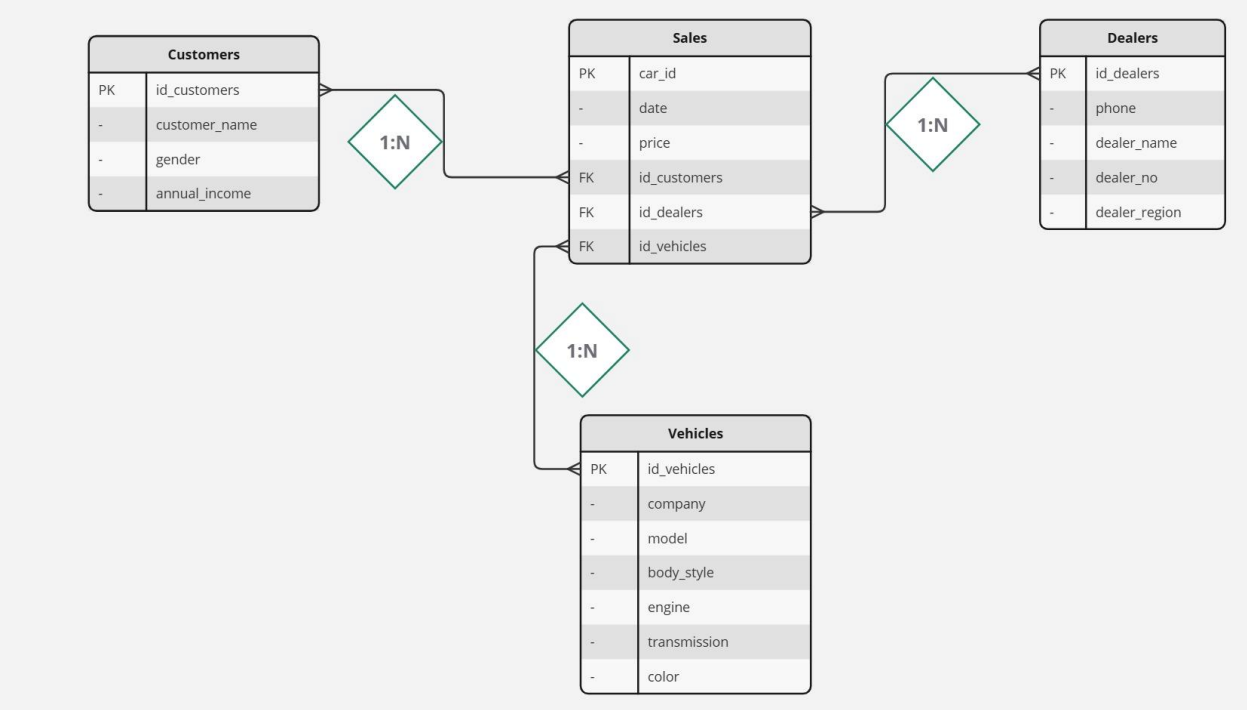


Figure 1: Entity-Relationship Diagram of the Database

List of Tables and Their Columns

Each table used in the entity-relationship model is described below, including the key definitions, the fields contained in each table, and the data types associated with those fields. The ERD consists of four main tables:

- Customers table
- Dealers table
- Vehicles table
- Sales table

Customers table		
Key	**Feature**	**Data types**
PK	id_customers	Int
-	customer_name	varchar(50)

-	Gender	varchar(6)
-	annual_income	Int

Table 2: Customers Table

The Customers table contains personal information about each client who made at least one vehicle purchase.

- The table includes the field id_customers as the primary key, which serves as the unique identifier for each customer.
- The customer_name field stores the name of the customer.
- The gender field represents the customer's gender.
- The annual_income field indicates the customer's annual income.

Dealers table		
Key	**Feature**	**Data types**
PK	id_dealers	Int
-	Pone	Int
-	dealer_name	varchar(100)
-	dealer_no	varchar(20)
-	dealer_region	varchar(20)

Table 3: Dealers Table

The Dealers table contains business information for each dealership that sold at least one car.

- The table includes the field id_dealers as the primary key, which serves as the unique identifier for each dealership.
- The phone field stores the phone number of the dealership.
- The dealer_name field represents the name of the dealership.
- The dealer_no field indicates the dealership number.
- The dealer_region field refers to the region where the dealership is located.

Vehicles table		
Key	**Feature**	**Data types**
PK	id_vehicles	Int
-	Company	varchar(20)
-	Model	varchar(25)
-	body_style	varchar(20)
-	Engine	varchar(50)
-	Transmision	varchar(6)
-	Color	varchar(20)

Table 4: Vehicles Table

The **Vehicles** table contains descriptive information about each of the vehicles available for purchase.

- The table includes the field id_vehicles as the primary key, which serves as the unique identifier for each vehicle.
- The company field stores the name of the vehicle's brand or manufacturer.
- The model field represents the model name of the car associated with the brand.
- The body_style field refers to the vehicle's body type.
- The engine field indicates the type of engine.
- The transmission field describes the vehicle's transmission type.
- The color field indicates the color of the vehicle.

Sales Table		
Key	**Feature**	**Data types**
PK	car_id	varchar(20)
-	Date	Date
-	Price	Int
FK	id_customers	Int
FK	id_dealers	Int
FK	id_vehicles	Int

Table 5: Sales table or Fact table

The Sales table is the fact table that records all car sales transactions.

- The table includes the field car_id as the primary key, which uniquely identifies each sales transaction.
- The date field indicates the date on which the sale was made.
- The price field represents the sale price of the vehicle.
- The id_customers field is the first foreign key, referencing the unique identifier of the customer involved in the sale.
- The id_dealers field is the second foreign key, referencing the unique identifier of the dealership that carried out the sale.
- The id_vehicles field is the third foreign key, referencing the unique identifier of the vehicle sold.

Relationship between tables

It is important to highlight the relationships that exist between the tables, specifically the connections between the primary keys in the Customers, Dealers, and Vehicles tables and their corresponding foreign keys in the Sales table. These relationships are described as follows:

- In the Customers table, the primary key id_customers directly relates to the foreign key id_customers in the Sales table.
- In the Dealers table, the primary key id_dealers corresponds to the foreign key id_dealers in the Sales table.
- Finally, the Vehicles table contains the primary key id_vehicles, which links to the foreign key id_vehicles in the Sales table.

Transformations performed

As part of the data modeling process, several transformations were applied, starting with the creation of a date (calendar) table based on the fact table (Sales). The first and last dates of vehicle sales transactions were identified, and this range was used in MS Excel to generate a list of all dates within that period. This list was then loaded into Power BI to serve as the project's calendar table, which was linked to the data model.

In addition to the calendar table, transformations were applied to each of the model's core tables:

Transformations in the Calendar table

- The column was renamed to Date.
- Null records were removed.
- The data type of the Date column was set to Date.

Transformations in the Dealers Table

- The first row of data was promoted to headers.
- The Phone column's data type was set to Whole Number.
- The Dealer_Region column was set to Text.
- The id_dealers column was set to Whole Number.

Transformaciones in the Customers Table

- The first row of data was promoted to headers.
- The Customer Name column was set to Text.
- The Gender column was set to Text.
- The Annual Income column was set to Whole Number.
- The id_customers column was set to Whole Number.
- In the Gender column, the values "Male" and "Female" were replaced with "Masculine" and "Femenine" respectively.

Transformaciones in the Vehicle table

- The first row of data was promoted to headers.
- Null records were removed from the table.
- The data type of the Company column was set to Text.
- The data type of the Body Style column was set to Text.
- The data type of the Engine column was set to Text.
- The data type of the Transmission column was set to Text.
- The data type of the Color column was set to Text.
- The data type of the id_vehicles column was set to Whole Number.

Transformations in the Fact Table (Sales)

- The first row of data was promoted to headers.
- The data type of the id_customers column was set to Whole Number.
- The data type of the car_id column was set to Text.
- The data type of the Price ($) column was set to Whole Number.
- The data type of the id_vehicles column was set to Whole Number.

- The data type of the id_dealers column was set to Whole Number.

Visualizations

The content of the Power BI dashboard and its respective visualizations is described in the following sections.

Objetive

The objective of visualizing the dataset using a specialized tool like Power BI is to enhance the understanding of the data, transforming raw data into actionable information and insights. This, in turn, facilitates knowledge transfer to the management team and supports data-driven decision-making based on objective numerical evidence.

Scope

The dashboard is designed for tactical use by branch leaders, enabling them to understand their sales performance and key figures. Additionally, it is intended to be presented at the strategic level to regional managers of the dealerships, allowing them to identify sales trends, make informed decisions, and develop tailored commercial strategies for different socio-economic groups within each region.

Calculated Measures

The calculated measures were created and organized within the **Measures Table**, and include the following:

- **Previous Period Sales:**

```
PreviousYearToDate = TOTALYTD(SUM('Tabla de hechos'[Price ($)]),
SAMEPERIODLASTYEAR('Tabla calendario'[Date].[Date]))
```

- **Average Accumulated Sales:**

```
Average Accumulated Sales = TOTALYTD(AVERAGE('Tabla de hechos'[Price ($)]), 'Tabla
calendario'[Date].[Date])
```

- **Total Sales Sum:**

```
Total Sales Sum = SUM('Tabla de hechos'[Price ($)])
```

- **Accumulated Sales:**

```
Accumulated Sales = TOTALYTD(SUM('Tabla de hechos'[Price ($)]), 'Tabla
calendario'[Date].[Date])
```

- **Sales Percentage Variation:**

```
YTD variation = ([Venta acumulada]/[PreviousYearToDate]-1)*100
```

Dashboard tabs

Color Palette

The following colors were used in the presentation:

- #FFFFFF
- #203864
- #000000
- #118DFF
- #F0F0E1
- #D64550

Cove Page

Figure 2: Cover Page

The cover page is the introductory section that displays the dashboard title—giving an initial idea of its theme—along with the owner's name, the organization's logo, and a navigation index to access the different dashboard tabs.

Glossary

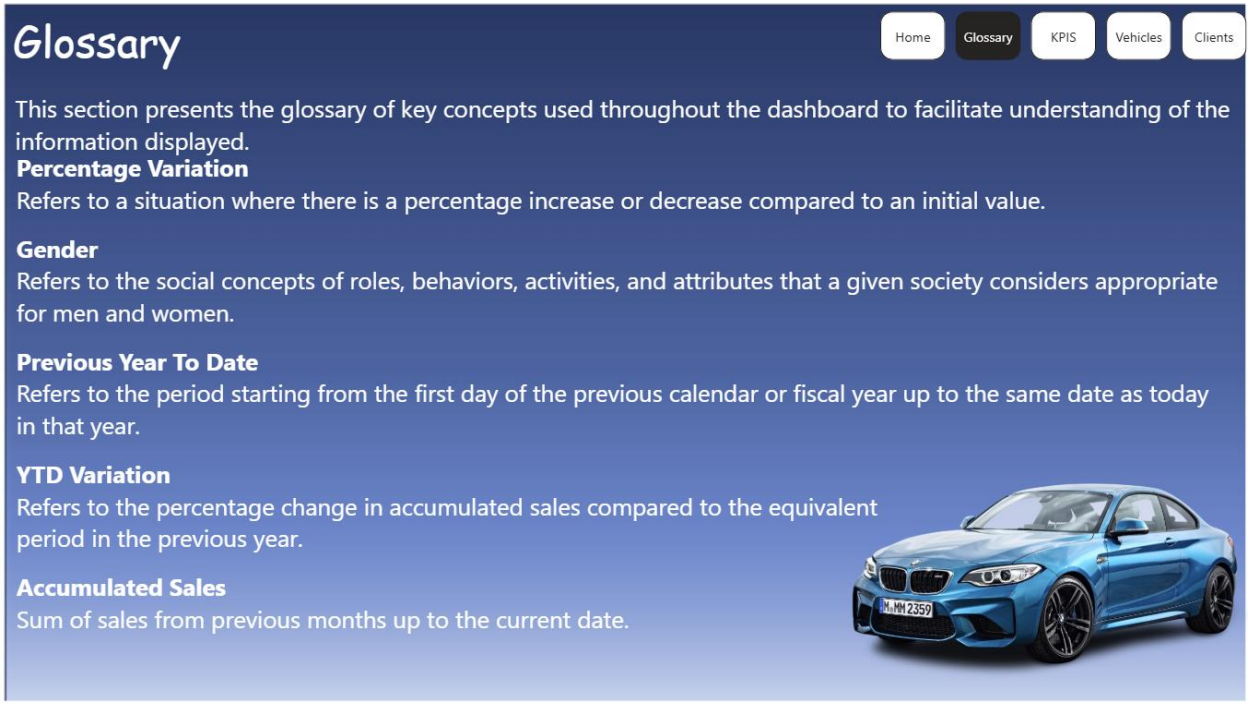


Figure 3: Glossary

The glossary presents the key concepts used throughout the dashboard, so that when we encounter them in the following tabs, we can clearly understand what they mean.

KPIS

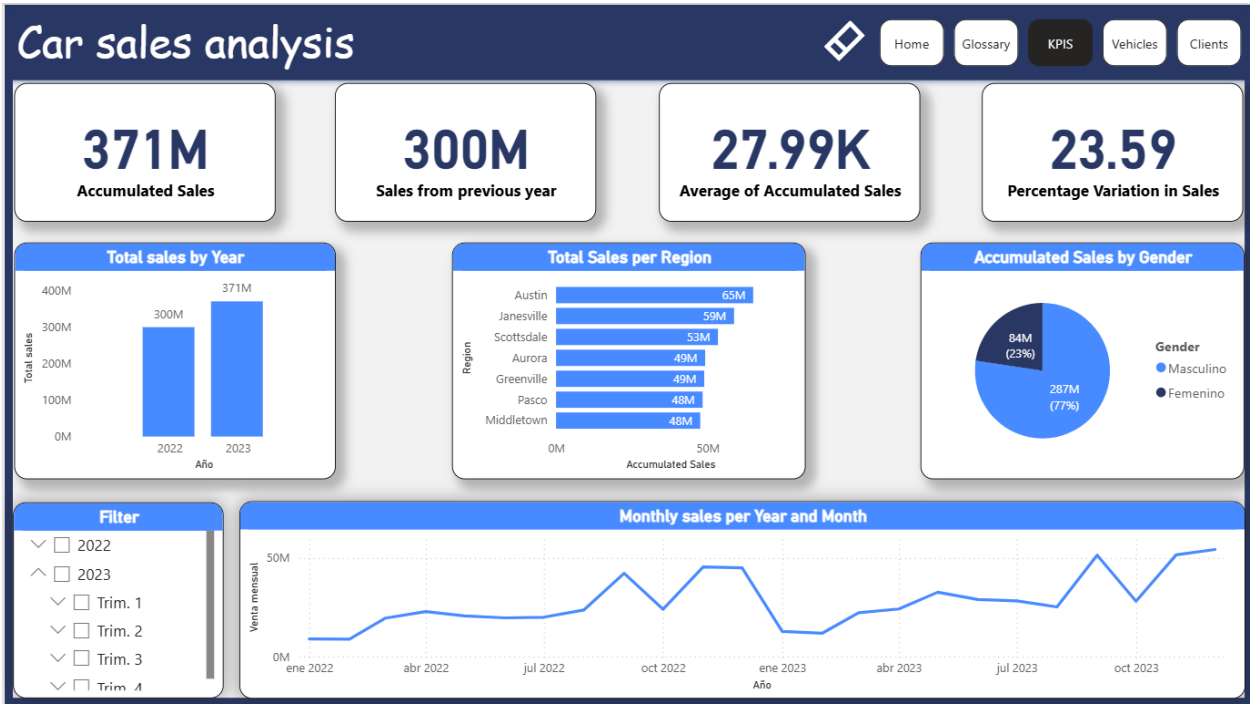


Figure 4: KPIS

The **KPIs tab**, titled *"Car Sales Analysis"*, displays various key performance indicators and visualizations related to car sales performance, as suggested by its title. This section includes accumulated sales, comparisons with the previous period, average sales, and percentage variation.

Additionally, it features charts that present sales by region, gender, and time period, as well as a time filter. This time filter enables users to perform different types of analysis based on the selected time frame.

This tab serves as the main dashboard page, offering high-level insights for management and decision-makers. It allows for the comparison of a current sales period with its equivalent in the previous year, helping to assess whether average sales are stable or shifting over time.

The visualizations support:

- Comparing sales differences across different time periods and regions,
- Understanding which gender group purchases more vehicles,
- Identifying monthly and yearly sales trends—useful for determining the times of year when the company generates the most sales.

Lastly, the time filter allows users to conduct the same analysis at more specific levels, such as quarterly, monthly, or even daily. This makes it possible to:

- Compare quarterly or monthly sales averages and assess whether performance is improving or declining,
- Analyze in which quarters or months female customers might be purchasing more than male customers,
- Evaluate gender-based and regional sales performance on a monthly basis.

Vehicles

Figure 5: Vehicles

The Vehicles tab provides a detailed analysis of car sales, breaking down the data to show how sales perform by vehicle brand. It also displays accumulated sales, as well as sales by vehicle color and body style.

These charts are interactive and can be adjusted using filters to show sales data by region and specific time periods.

By using filters for region, time, and gender, users can:

- Analyze how the top 10 best-selling brands change based on selected criteria.
- Understand which vehicle styles and colors are sold most frequently across different regions, brands, genders, and time periods.

Customers

Figure 6: Customres

The Customers tab provides insight into sales behavior at the customer level. It includes visualizations showing the average income by gender and how sales have evolved over time by gender. Additionally, it presents accumulated sales segmented by gender and region, as well as by vehicle style and gender.

This type of information helps identify customer income levels, allowing the company to tailor its vehicle pricing strategies accordingly. It also highlights sales peaks by gender and indicates which gender group purchases more vehicles.

Entity-Relationship Diagram in Power BI

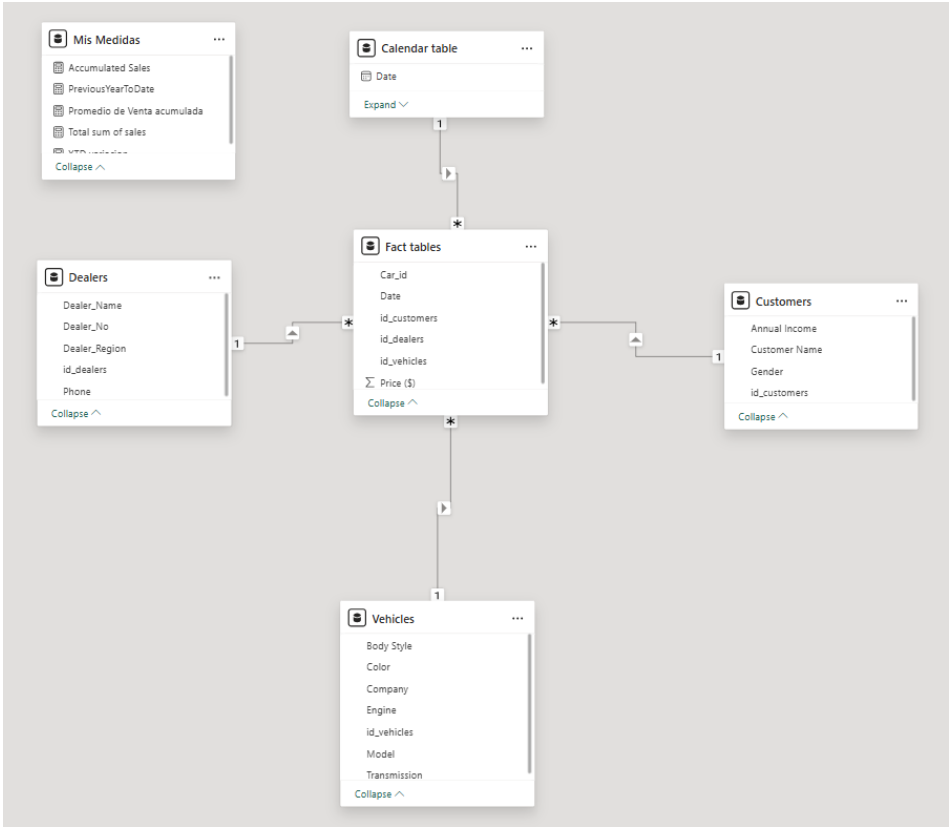
1. 

Figure 7: Entity-Relationship Diagram in Power BI

The entity-relationship diagram (ERD) built in Power BI was based on the original ERD of the underlying database. The main difference between this diagram and the original database ERD is that the Power BI version includes two additional tables: the Calendar Table and the Measures Table (or “My Measures”).

The fact table, or Sales table, contains all vehicle sales transactions. This table includes the fields Date, id_customers, id_dealers, and id_vehicles, which serve as foreign keys linking to the Dealers, Customers, and Vehicles tables, respectively.

The Customers table contains customer-related information, such as annual income, customer name, and gender. It includes the primary key id_customers, which connects to the corresponding foreign key in the Sales table.

The Dealers table holds data related to the dealerships, including dealer name, region, phone number, and the primary key id_dealers, which links to the foreign key id_dealers in the Sales table.

The Vehicles table stores detailed vehicle information, such as body style, color, brand, engine type, model, and transmission type (manual or automatic). This table’s primary key, id_vehicles, links to the foreign key id_vehicles in the Sales table.

As mentioned earlier, the Calendar Table is not present in the original database ERD but was added in Power BI to support time-based analysis. This table contains only the Date column, which links to the Date column in the Sales table.

Lastly, the Measures Table stores all calculated measures. It is not connected to any other table in the model.

Futures Initiatives

From a deeper analysis of the information displayed in the Vehicles tab, we can develop branch-level strategies, such as increasing the inventory of pale white cars, which were identified as the top-selling color, while reducing the stock of less popular colors like red. At a regional or branch level, we can use these insights to align inventory with customer preferences by region, gender, vehicle brand, style, and color, ensuring higher availability of the vehicles customers truly want and reducing those with lower demand.

In the Customers tab, knowing the breakdown of sales by gender allows us to shape targeted acquisition strategies, especially to appeal to female customers. This could involve price-based promotions or introducing new color options. To support such product launch strategies, collaboration with the marketing team is essential, they should conduct surveys to determine which vehicle colors are most attractive to female customers in each region. With this information, dealerships can stock vehicles aligned with these preferences, ultimately increasing sales among this customer segment.

Additionally, the dashboard reveals that the first two months of the year consistently show the lowest sales volume regardless of gender. This insight suggests the opportunity to implement a clearance strategy for older vehicle models during this slow season, offering them at discounted prices to boost sales in an otherwise low-performing period.

Another future initiative worth considering is analyzing customer preferences for transmission type, automatic vs. manual. This preference may vary by both gender and region, and could be valuable for further personalizing inventory and marketing strategies.

Finally, a potential long-term commercial strategy would be the introduction of electric vehicles (EVs). However, this initiative would require marketing-led surveys to assess customer interest in EVs. Analyzing the results could help the company offer selected electric vehicle models and gradually expand into this market segment while minimizing risk and exploring new opportunities.

Conclusions

At the beginning of this project, we hypothesized that all car brands might sell similarly across different regions and socio-economic groups. However, after conducting a rigorous analysis of the car sales dataset, examining variables such as income, gender, individual preferences, and the availability of specific models, brands, and colors, we found that this hypothesis must be rejected. Sales behavior is not uniform. Each region and socio-economic segment exhibits distinct preferences and purchasing patterns.

This project has revealed valuable insights into consumer behavior. We identified clear preferences in terms of car brand, vehicle style, and color, which also vary by region. These findings have enabled the identification of specific customer segments with similar consumption habits. This segmentation, in turn, allows the company to offer more tailored products to meet the unique needs of each target group more effectively.

Data analysis is a continuous process. It should not be limited to this dataset alone, ongoing monitoring is necessary to evaluate whether trends persist or shift in future periods. Additionally, new datasets

could be analyzed, such as those containing more detailed customer information, to conduct a deeper customer analysis that complements and enhances the insights generated in this project.